Rudolf Schmidt

WÖRTERBUCHARTIKEL UND ERGEBNISDATENBANK

Wie allgemein bekannt sein dürfte, beabsichtigt das Institut für deutsche Sprache die beiden in Bearbeitung befindlichen Lexika 'Lexikon der schweren Wörter' und 'Lexikon der Lehnwortbildung' computerunterstützt zu erstellen. Es soll deshalb im folgenden kurz ein Einblick in die dazu gemachten Ideen und Vorstellungen gegeben werden, wobei der Schwerpunkt der Ausführungen auf den Strukturen von Wörterbuchartikeln und den ihnen zugeordneten Datenmodellen liegt.

Zunächst sollen die Komponenten einer Lexikographischen Datenbank, wie man sie am IdS vorstellt, aufgezeigt werden. Anschließend wird die Abbildbarkeit von Artikelstrukturen auf Daten(bank)strukturen dargestellt. Im Anschluß daran werden die Vorteile einer Lexikographischen Datenbank für den Anwender gezeigt, wobei jeweils die Bedeutung der Artikelstruktur hervorgehoben wird. Schließlich soll eine Möglichkeit der Realisierung vorgestellt werden.

1. Komponenten einer Lexikographischen Datenbank

Nach Auffassung sowohl der Lexikographen wie der Mitarbeiter der LDV des IdS sollte eine Lexikographische Datenbank, kurz LEDA genannt, im großen und ganzen aus folgenden Komponenten bestehen:

1. einer Textdatenbank zur Belegsuche
2. einer Wortdatenbank, die Lemmalisten enthält und darüber hinaus auch Lexeme in segmentierter Form, u.a.m.
3. einer bibliographischen Datenbank für eine einheitliche Zitierung sowohl der Belege aus 1., als auch sonstiger nicht in 1. abgelegter Belege
4. einer temporären Arbeitsdatei, in die der Lexikograph seine Artikel schreibt, und in der er sie bearbeitet
5. der Ergebnisdatenbank, in die die Artikel aus der Arbeitsdatei nach Fertigstellung durch den Lexikographen und Durchlaufen bestimmter Prüfroutinen kopiert werden

Die Komponenten 'Textdatenbank' und 'Bibliographische Datenbank' sind weitgehend schon realisiert und bedürfen im Zusammenhang mit LEDA höchstens einiger Anpassungen. Auch eine 'Wortdatenbank' ist in Form der am 'Institut für Kommunikationsforschung und Phonetik' Bonn erstellten 'kumulierten Wortdatenbank' vorhanden. Eine Integration dieser Komponente in das Gesamtsystem ist im Augenblick zurückgestellt, da die dort enthaltenen Angaben ausschließlich morpho-syntaktischer Art sind. Diese Angaben sind für die derzeit in Bearbeitung befindlichen Lexikon-Projekte von untergeordneter Bedeutung. Zudem verfügt diese 'Wortdatenbank' nicht über segmentierte Wortformen, die, wie wir später sehen werden, von Anfang an von Bedeutung sind.

Die Trennung zwischen Arbeitsdatei und Ergebnisdatenbank ist aus Gründen der Datensicherheit und der Konsistenz vorgesehen.

2. Abbildbarkeit von Artikelstrukturen auf Datenbankstrukturen

Im folgenden soll gezeigt werden, welche Beziehungen zwischen den Strukturen eines Artikels und den Strukturen einer Datenbank bestehen, welche Bedeutung dies hat für

1. den Zugriff auf Teile des Lexikons
2. das Layout
3. (a) die Prüfung der Konsistenz des Wörterbuchs
   (b) die Konsistenzerhaltung bei Umänderungen innerhalb eines Artikels z.B. Umsortierung von Unterartikeln

Um Artikel- und Datenbank-Strukturen besser verständlich zu machen, seien an dieser Stelle kurze Erläuterungen und Definitionen erlaubt.

In der Computerlinguistik besteht ein Wörterbuch zunächst aus einer Menge von lexikalischen Einheiten, auch Artikel genannt,

$$W = \{ LE_1, \ldots, LE_n \}$$

wobei die lexikalischen Einheiten aus Identifikations- und Informationsteil bestehen. Für manche Zwecke ist diese Unterteilung noch zu grob und man unterscheidet beispielsweise noch zwischen einzelnen Artikelpositionen, so z.B. zwischen Angaben zur Phonetik, Morphologie, Syntax, Semantik, Etymologie oder zum Gebrauch. Dies läßt sich formal folgendermaßen darstellen:

$$LE_i = (I_{i1}, I_{i2}, \ldots, I_{in})$$

wobei I_i die i-te Informationseinheit darstellt.

Das heißt, daß sich ein Artikel als (geordnetes) Tupel schreiben läßt. Eine ähnliche Strukturierung liegt auch der Theorie des relationalen Datenbankmodells zugrunde. Hier werden die logisch zusammengehörigen Komponenten in einer Datei in Form einer Tabelle abgespeichert. Die Zeilen stellen die Objekte, auch Entities oder Sätze genannt, die Spalten die Attribute, auch Felder genannt, dar.

In dieses Schema passen vor allen Dingen streng strukturierte Angaben, wie z.B. morpho-syntaktische Informationen oder Wörterbuchartikel, wie z.B. der folgende:

Artikel 'Revanchismus'

[1] Revanchismus, [2] der; [3] -, nur Sing.

[4] wird vom Sprecher als politisches Schlagwort mit stark negativer Wertung verwendet, um den jeweiligen Gegner bloßzustellen und zu diffamieren

[5] mit **R.** charakterisiert der Sprecher eine politische Einstellung, Handlungsweise, Ideologie und eine darauf beruhende Politik, die er aufgrund ihrer Beweggründe als reaktionär und aggressiv einschätzt:

[6] 'Politik, die durch Rachegelüste und Vergeltungsabsichten des Besiegten gegenüber dem Sieger nach einem Krieg gekennzeichnet und auf Veränderung der durch ihn geschaffenen Verhältnisse ausgerichtet ist, insbesondere die Annullierung von (vermeintlich) aufgezwungenen Verträgen und die Rückgewinnung verlorener Gebiete mit militärischen Mitteln anstrebt; Revanchepolitik, Revanchebestrebungen, Vergeltungspolitik';

[7] **R.** wird speziell als kommunistisches Kampfwort verwendet für die der BRD oder bestimmten politischen (z.B. nationalistischen) Kreisen in der BRD zugeschriebenen friedensfeindlichen Tendenzen, die Ergebnisse des 2. Weltkriegs in Europa auf militärischem Weg rückgängig zu machen, daher wird R. häufig in Verbindung mit anderen negativ wertenden Schlagwörtern wie *Militarismus, (Neo)Nazismus, Imperialismus, Chauvinismus, Antikommunismus* und in den festen Wendungen *der westdeutsche R., der Bonner R.* gebraucht. -

[8] Sinnverwandt: *Reaktion, Revisionismus*. Gegensatz: *Pazifismus*. -

[9] Beispiele: den R. vertreten; die Gegner des R.; der R. in der BRD; der R. als Form des übersteigerten Nationalismus. -

[10] Textvorkommen: Vorwiegend in wertenden Zeitungstexten (Berichten über die internationale Politik, in Kommentaren und Stellungnahmen). -

[11] ∆ Als politisches Schimpfwort wird R. taktisch oder manipulativ stets auf den weltanschaulichen Gegener bezogen!

[12] ↑ **SCHIMPF- und KAMPFWÖRTER, POLITISCHE. ISMEN, POLITISCHE.**

[13] * Ableitungen: *Revanchist ... revanchistisch ...*

[14] ◊ Etymologie: R. ist eine relativ junge Ableitung (ca. seit 1954 nachweisbar) von wesentlich älterem ↑ *Revanche;* vgl. dagegen *revanchistisch,* das vereinzelt schon im frühen 20. Jh. (1919) belegt ist.

Verwendungsgeschichte: Besonders häufig wird R. in den 50er und 60er Jahren von der DDR auf die Politik der BRD angewandt, die die Abtretung der deutschen Ostgebiete jenseits der Oder-Neiße-Linie an Polen entsprechend den Bestimmungen des Potsdamer Abkommens von 1945 nur als vorläufig betrachtete. Erst seit der Unterzeichnung des Vertrags zwischen Polen und der BRD von 1970, in dem diese die Oder-Neiße-Linie als Westgrenze Polens faktisch anerkannte, und des Grundlagenvertrags zwischen der BRD und der DDR von 1972 ist der Gebrauch von R. in der deutsch-deutschen Politik rückläufig.

[15] Belege:
zugleich aber ... wird uns mit dem Buch eine scharfe Waffe für unseren heutigen Kampf um die Herstellung eines geeinten demokratischen Deutschlands, gegen den westdeutschen Militarismus, den Revanchismus und die Kriegsvorbereitungen des amerikanischen Imperialismus und seiner westdeutschen Verbündeten in die Hand gegeben (Neues Deutschland 4.4.1954); jeder Revanchismus gegenüber den osteuropäischen Völkern ist abzulehnen, der Rechtsanspruch auf die deutschen Ostgebiete darf jedoch nicht preisgegeben werden (Die Welt 30.11. 1959); Revanche heißt wörtlich übersetzt etwa "Vergeltung". Der Begriff Revanchismus, der nicht besonders glücklich von "Revanche" abgeleitet wurde, bezeichnet im politischen Sprachgebrauch eine politische Einstellung oder Handlung, der als Beweggrund ... Vergeltungsideen zugrunde liegen (Süddeutsche Zeitung 2.7.1960); Seit Anfang 1966 fehlen in Chinas Presse und Rundfunk ... die ... üblichen Schmähungen gegen westdeutschen Revanchismus, Kriegstreiberei und ähnliches (Stuttgarter Zeitung 30.1.1968); durch die Bekämpfung von Kriegshetze, Revanchismus, Nazipropaganda, Rassen- und Völkerhaß eine stabile Friedensordnung in der Deutschen Demokratischen Republik herbeizuführen (Neues Deutschland 8.12.1969); so hat Brandt ... die ostdeutschen Behauptungen durchlöchert, wonach einer Lösung des deutschen Problems nur westdeutscher Revanchismus, Militarismus und Neonazismus entgegenstehen sollten (FAZ 16.2.1970); während Kohl der SPD vorwarf, kommunistischen Propagandisten Stichworte für eine "absurde Revanchismuskampagne" zu liefern (Mannheimer Morgen 29.1. 1985); zugleich wies der Parteitag [der FDP] aber auch "ungerechtfertigte Revanchismusvorwürfe" des Ostens gegen die Bundesrepublik zurück (Mannheimer Morgen 25.3.1985).

[16] Literatur: ...

Die in eckigen Klammern stehenden Ziffern gehören dabei nicht zum Artikel sondern dienen lediglich der Bezugnahme auf die sie repräsentierenden Textbausteine bzw. Artikelpositionen. Dabei bezeichnet die Position 1 das Lemma, Position 2 die Wortartbestimmung (z.B. Genus), 3 = flexionsmorphologische Angabe, 4 = Funktionsklasse (Kommunikationsbereich und Wertungsdimension), 5= Funktionsbeschreibung, 6 = präzisierender Teil der Bedeutungserläuterung, 7 = spezieller Gebrauch, 8 = sinnverwandte lexikalische

Strukturbeziehung, 9 = Syntax, 10 = Textvorkommen, 11 = sprachkritische Vermerke, 12 = Verweis von|für 11, 13 = Zusatzinformation, 14 = Etymologie, 15 = Textbelege, 16 = Literatur.

Wichtig in diesem Zusammenhang ist, daß diese Positionen - es handelt sich dabei um die des 'Lexikons der schweren Wörter' - so weit gefaßt sind, daß sie in allen Artikelsorten des Lexikons enthalten sein können, d.h. sowohl bei Einzel- als auch bei Rahmenartikel (vgl. Vortrag Gerhard Strauß).

Der oben genannte Artikel besteht also aus Textbausteinen, die

1. eine eindeutige Funktion besitzen
2. in aufsteigender Reihenfolge stehen
3. alle in diesem Artikel vorkommen (d.h. alle Positionen sind nicht leer)

Es fällt weiterhin auf, daß der Umfang der einzelnen Artikelpositionen sehr stark schwankt (vgl. z.B. 1, 2, 3, 8, 13 mit 6, 7, 14). Bezüglich der Realisierung mittels eines kommerziellen relationalen Datenbanksystems stellt hier vor allen Dingen die maximale Feldlänge eine Grenze dar. Diese liegt bei Systemen, die auf der SIEMENS-Anlage des IdS implementiert sind, bei 255 Zeichen (SESAM), 965 Zeichen (FIDAS) und über 16.000 Zeichen (PINDAR).

Es gibt natürlich auch Systeme, leider jedoch nicht auf der Anlage des IdS, die eine Feldlänge von über 65.000 Zeichen erlauben, z.B. KNOWLEDGEMAN und ORACLE unter MS-DOS. Eine derartige Begrenzung dürfte eigentlich keine Einschränkung mehr für die Lexikographie bedeuten, wenn man bedenkt, daß ein Artikel aus mehreren Textbausteinen bzw. Artikelpositionen zusammengesetzt ist.

Eine Verteilung der Artikelpositionen auf mehrere 'Felder', wie diese Einheiten in der Notation der Datenbanken bezeichnet werden, ist nicht wünschenswert, weil dadurch eine eindeutige Zuordnung von 'Feld' und 'Artikelposition' nicht mehr gewährleistet ist, und die logische Struktur damit undurchsichtig wird. Um Inkonsistenzen bei der Abfrage zu vermeiden, müßten die Ausgabeprogramme u.U. komplexe Fallunterscheidungen treffen, um die eigentliche Struktur zu rekonstruieren, ein Umstand, der den Einsatz einer (konventionellen) Datenbank nicht mehr rechtfertigt.

Außer dieser Beschränkung gibt es noch weitere technische Hindernisse. So ist es beispielsweise mit FIDAS oder PINDAR nur möglich, auf maximal zwei Dateien gleichzeitig zuzugreifen. Das Schlüsselfeld (Feld, das das Objekt identifiziert und über das i.d.R. ein schneller Zugriff erfolgen kann), darf z.B. bei FIDAS die Länge 8 (!) nicht überschreiten. Damit ist es nicht möglich, Artikel über das Lexem zu identifizieren.

Beschränkt man sich zudem nicht wie oben auf Artikel mit einer Struktur, bei der die Informationseinheiten I_{il}, ... , I_{im} klar voneinander abgegrenzt und geordnet sind, so treten weitere Schwierigkeiten bei der Datenmodellierung auf. Betrachtet man z.B. einen Artikel, der im erzählerischen Stil geschrieben ist, so kommen folgende Fälle oft vor:

1. Teile des Artikels lassen sich u.U. mehreren sogenannten Artikelpositionen zuordnen
2. Artikelpositionen können alternierend auftreten, z.B. 6a, 7a, 6b, 7b u.dgl.
3. Artikelpositionen können geschachtelt sein, z.B. 6a, 7a, 8, 7b, 6b u.dgl.

Bei der oben vorgestellten Wörterbuchstruktur lassen sich nur die Positionen 1, 2, 3, 8 und 14 weitgehend standardisieren. Andere Artikelpositionen, insbesondere 4, 5, 6 und 7, die unter der Bezeichnung 'handlungssemantischer Kommentar' zusammengefaßt werden, können eher als weniger standardisiert, d.h. diskursiv angesehen werden.

Dies soll der folgende Artikel illustrieren:

Artikel 'Metapher'

[1] Metapher, [2] die; [3] pl. -n; adj. metaphorisch

[4] Bezeichnet eine bestimmte Weise der Verwendung sprachlicher Ausdrücke, die meist als nicht wörtlich oder übertragen charakterisiert wird.

[7a] Nehmen wir die beiden folgenden Äußerungen, in denen das Wort Fuchs zunächst nicht metaphorisch und dann metaphorisch verwendet ist. ...

[6a] Wir können jetzt die Charakterisierung metaphorischer Sprachverwendung als nicht wörtlich oder übertragen so präzisieren: ...

[7b] Diese Präzisierung bringt uns auf die folgende Frage: wodurch unterscheiden sich Äußerungen mit metaphorischer Sprachverwendung wie 'dieser Junge ist ein Fuchs' von Äußerungen ohne metaphorische Sprachverwendung wie 'dieser Junge ist schlau/listig'? ...

[6b] Äußerungen mit metaphorischer Sprachverwendung unterscheiden sich von Äußerungen ohne metaphorischer Sprachverwendung dadurch, daß sie die besonderen Eigenschaften oder Dispositionen, die mit ihnen jeweils ausgedrückt werden, unter einer bestimmten Sichtweise vermitteln, ...

[5] Sprecher, die sich metaphorischer Ausdrucksweisen bedienen, versuchen damit auch immer - mehr oder weniger bewußt -, ihre Adressaten zu Komplizen ihrer Sichtweisen zu machen. ...

[7c] Die metaphorische Verwendung eines sprachlichen Ausdrucks ist nicht an eine bestimmte Wortart oder an eine bestimmte Funktion im Satz gebunden, ...

[11] Eine kritische Haltung gerade gegenüber metaphorischen Ausdrücken wie 'Ratten', 'Schmeißfliegen', 'Ungeziefer' u.ä. ist besonders nötig, denn Sprecher, die sich dieser Ausdrücke bedienen, stellen sich in eine Tradition der Rede- und Sichtweise, die in der Zeit des Nationalsozialismus mit Bezug auf Juden und Intellektuelle üblich war. ...

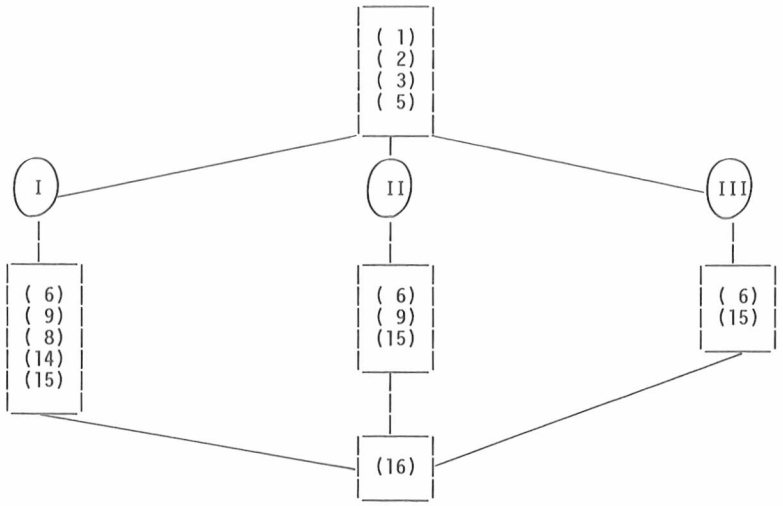
[10] Textvorkommen

Bei den bisherigen Ausführungen haben wir noch völlig außer acht gelassen, daß ein Artikel über eine 'hierarchische' Struktur verfügen kann, d.h. ein solcher Artikel ist gegliedert in eine oder mehrere Hauptbedeutungen des Lemmas, wobei jede Hauptbedeutung eine oder mehrere Unterbedeutungen aufweist, die wieder unterteilt sind in eine oder mehrere Unter-Unterbedeutungen. Hierbei werden dann die unterschiedlichen Bedeutungsebenen parallel zueinander erklärt.

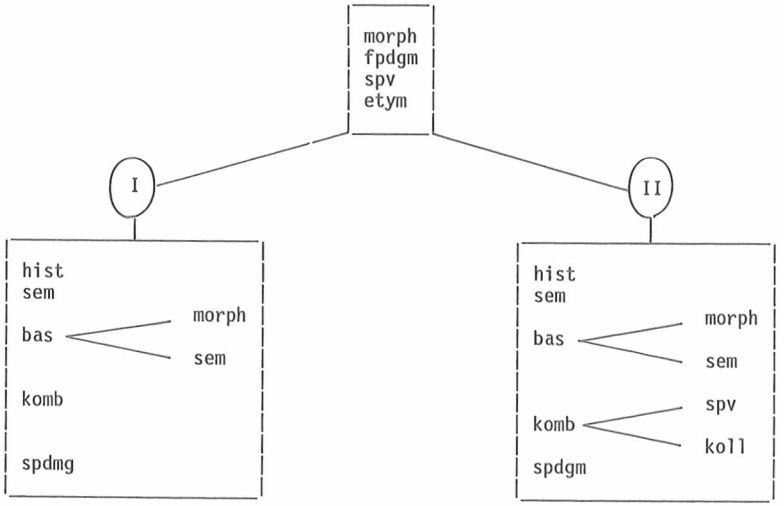
Auf den beiden folgenden Seiten ist die Struktur von drei verschiedenen Artikeln dargestellt. Es handelt sich um die Lemmata *Entsorgung* ('Lexikon der schweren Wörter'), *-itis* ('Lexikon der Lehnwortbildung') und *da* (sechsbändiger DUDEN).

Steht man vor dem Problem, eine solche Struktur auf ein Datenmodell abzubilden, so greift man intuitiv zu dem hierarchischen Datenmodell. Dieses strukturiert, wie der Name schon sagt, Datenobjekte und ihre Beziehungen als Baumstrukturen, d.h. übergeordnete Segmente (bzw. Objekttypen) können viele untergeordnete Segmente besitzen, während untergeordnete Segmente genau ein übergeordnetes Segment haben. Die Über- und Unterordnung von Segmenten kann bis zu einer vorher im Schema festgelegten Tiefe geschachtelt werden. Damit lassen sich baumartig strukturierte Datenobjekte mit fester Struktur direkt modellieren und effizient implementieren. Weniger geeignet ist dieses Datenmodell für solche Anwendungen, deren hierarchische Struktur sich nicht von vornherein auf eine Baumstruktur bestimmter Tiefe festlegen läßt, deren Objektkomponenten variabel groß sein können

Struktur des Artikels 'Entsorgung'

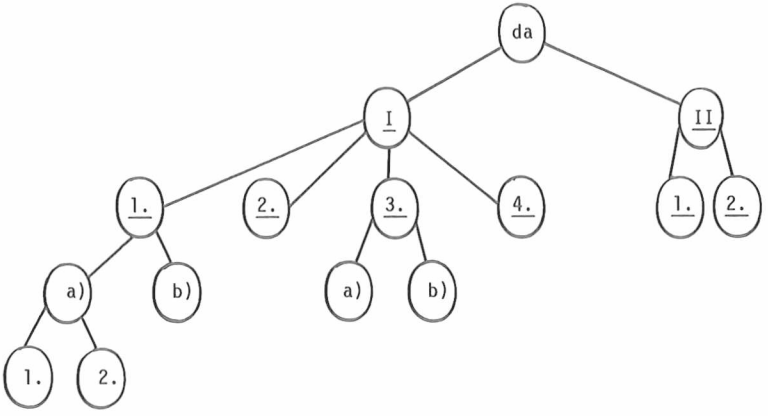


Struktur von '-itis'



Graphische wie textuelle Darstellung des Artikels 'da' aus dem sechsbändigen DUDEN

An diesem Beispiel erkennt man sehr deutlich, daß der die Hierarchisierung symbolisierende Baum sehr ungleichmäßig sein kann.



da [da:; I, 1, 3: mhd. dä(r), ahd. där; I, 2: mhd., ahd. dö]: (I) < Adv. > (1.) (räumlich; hinweisend) a) *an dieser Stelle, dort:* da vorn; er wohnt da; * da und da *(irgendwo, an einem nicht näher bezeichneten Ort);* da und dort (1. *an einigen Orten, an manchen Stellen.* 2. *manchmal, hin und wieder);* b) *hier:* da sind wir; da nimm das Geld! (2.) (zeitlich) *zu diesem Zeitpunkt, in diesem Augenblick:* da lachte er; von da an herrschte Ruhe. (3.) (modal) a) *unter diesen Umständen, unter dieser Bedingung:* wenn ich schon gehen muß, da gehe ich lieber gleich; b) *in dieser Hinsicht:* Wenn ich Ihnen da einmal etwas zeigen darf, mein Herr, wir haben gerade einige neue Muster (von Verlobungsringen) hereinbekommen (Kant, Impressum 202). (4.) < als Teil eines Pronominaladverbs in getrennter Stellung > : ↑ dabei (5), dafür (7), dagegen (6), daher (4), damit (2), danach (4), dazu (4). (II) < Konj. > (1.) (begründend) *weil:* da er krank war, konnte er nicht kommen; Ich gebe meine Erinnerungen mit aller Vorsicht wieder, da ich mich auf manche Einzelheit nicht mehr genau besinnen kann (Jens, Mann 155); < mit vorausgehender Zeitbestimmung: > jetzt, da feststeht, daß die Wiedervereinigung ... nur mit Hilfe der Westmächte zu erreichen ist (Dönhoff, Ära 218). (2.) (zeitlich: geh.) *als:* da er noch reich war, hatte er viele Freunde; < mit vorausgehender Zeitbestimmung: > Die Erde war zu der Zeit, da man sie für eine Scheibe hielt, gewiß nicht weniger rund als heute (Dönhoff, Ära 108).

(vgl. *Revanchismus)* oder die nicht von Natur aus hierarchisch strukturiert sind.
Berücksichtigt man, daß sich einzelne Artikelpositionen auf verschiedenen Hierarchieebenen aufeinander beziehen können, so reicht das hierarchische Modell nicht mehr aus: man erhält ein sogenanntes Netz. Auch hierfür gibt es ein Datenmodell, das sogenannte Codasyl-Modell, das es erlaubt, netzwerkartige Strukturen abzubilden. Insbesondere wird die Einschränkung auf 1:1- und 1:n-Zuordnungen, wie sie beim hierarchischen Datenmodell vorliegt, erweitert durch die Möglichkeit von m:n-Abbildungen. Die Schwierigkeit, die beim Versuch auftaucht, Artikelstrukturen vermittels dieses Datenmodells darzustellen, ist, daß analog zum hierarchischen Datenmodell die Struktur der Artikel zu heterogen ist, um sich durch dieses zwar sehr allgemeine jedoch wenig flexible Datenmodell darstellen zu lassen.

3. Zugriffsmöglichkeiten für den Benutzer

Wozu, wird sich mancher fragen, braucht man ein Datenbanksystem zur Lexikonerstellung? Wozu der Aufwand, Artikelstrukturen auf Datenbank-Strukturen abzubilden? Reicht es nicht, Lexika einfach in maschinenlesbarer Form zu haben, um damit den Vorteil der leichten Änderbarkeit der Artikel und die Vorzüge eines Textverarbeitungssystems mit den vielfältigen Möglichkeiten der typographischen Gestaltung zu nutzen sowie den leichten Zugriff über Suchroutinen auf bestimmte ein- oder mehrwortige Suchbegriffe? Dazu ist zu sagen, daß zwar Wörterbücher der konventionellen Art m.E. nur über die Angabe des Lemmas Informationen zur Verfügung stellen, daß diese Einschränkung jedoch technisch bedingt ist. Hier liefern Konzepte aus der Datenbanktheorie wertvolle Hinweise zur Realisierung komplexerer Zugriffsmöglichkeiten. Diese können sowohl beim Artikelschreiben dem Lexikographen eine Hilfe sein, wenn sie beispielsweise ermöglichen, ganze Wortfelder im Zusammenhang darzustellen, als auch nach der Erstellung des Lexikons die Möglichkeit bieten, nur Teile auszudrucken, die nach bestimmten Kriterien selektiert wurden. Je genauer ein Wörterbuch strukturiert ist, und je unabhängiger die einzelnen Artikelpositionen voneinander sind, um so mehr Möglichkeiten der Selektion und damit der Nutzbarmachung der vorhandenen sog. 'virtuellen Lexika' bieten sich dem Benutzer.

So könnte man aus einem Wörterbuch mit umfangreicher Information, klare Strukturierung vorausgesetzt, ohne weiteres Wörterbücher zur Rechtschrei-

bung, Etymologie-, Synonym-, Stil- und eine Reihe von Fachwörterbüchern erhalten, ohne den Prozeß der Erstellung neu durchlaufen zu müssen.

Die Strukturierung muß jedoch nicht bei der Mikrostruktur des Wörterbuchs stehen bleiben. Auch eine andere Anordnung der Lexeme ist denkbar und kann u.U. aus didaktischen Gründen angebracht sein. So können z.B., wie dies bei Valenzwörterbüchern üblich ist, Lexeme nach semantischen Feldern sortiert sein.

Für die Lexikon-Projekte des IdS sollten diese erweiterten Möglichkeiten voll ausgeschöpft werden, sowohl für die Lexikonerstellung als auch -benutzung.

Als Beispiele für Abfragemöglichkeiten seien folgende genannt:

1. Zugriff auf Wort- und Morphemebene

   Eine Abfrage zu 1. könnte z.B. so aussehen, daß zunächst alle Wortformen gesucht werden, die das Morphem *Meta* enthalten. Ein Ausschnitt der vom Computer zur Verfügung gestellten Liste könnte etwa so aussehen:

   Metapher
   Metaphysik
   Metasprache
   usw.

   Jedoch nicht *Metallarbeiter, Metallindustrie* o.ä. Daher der oben genannte Wunsch, eine wortsegmentierte Lemmaliste in der Wortdatenbank zu haben!

2. Abfrage nach grammatischen, pragmatischen oder sonstigen Attributen:

   z.B. Liste aller Substantive im Femininum
   (Abfrage Position 2)

3. Direkter Zugriff auf eine ausgewählte Artikelposition:

   z.B. interessieren nur die Textbelege von *Revanchismus*
   (Position 14)

4. Kombinationssuche:

   z.B. Liste aller Substantive im Maskulinum,
   die nur im Singular gebräuchlich sind
   (z.B. *Revanchismus*, Position 2 + 3)

5. Abfrage nach dem Vorhandensein einer fakultativen Artikelposition:

z.B. Liste aller Lemmata mit sprachkritischem Vermerk (d.h.: Ist die Position 11 vorhanden?)

6. Abfrage, ob eine Wortform mehrere Bedeutungen hat (Polysemie):

z.B.: Sind bei dem Artikel *da* mehrere Bedeutungsvarianten vorhanden? Ergebnis: Gebräuchlich als Adverb und als Konjunktion (I u. II).

4. Typographie

Des weiteren kann man die logische Struktur eines Artikels eng mit seiner typographischen Gestaltung verbinden, d.h. man kann über das Layout die Strukturierung leichter erkennbar machen und hierdurch dem Benutzer eine weitere Erleichterung an die Hand geben. Liegt einmal die Struktur eines Artikels fest, so kann die dazugehörige Gestaltung des Layout beliebig gewählt werden, und zwar ist der Lexikograph von der Angabe bestimmter Steuerzeichen zur Gestaltung des Layout entbunden, da dies über die Position geschieht. Ob die Position des Artikels petit, *kursiv* oder anders gedruckt wird ist dann eine Entscheidung, die für das gesamte Lexikon global und nur einmal getroffen werden muß. Ähnlich leicht läßt sich festlegen, daß beispielsweise das Lemma in bestimmten Artikelpositionen *kursiv* oder **fett** gedruckt wird, in anderen Positionen normal oder als Abkürzung u.dgl.

Als Beispiel sei hier auf den oben genannten Artikel *Revanchismus* verwiesen. Hier wird das Lemma in den Positionen 5, 7, 9, 11 und 14 durch *R.* abgekürzt, in Position 15 voll ausgeschrieben. Ähnliches gilt auch für die anderen Artikel.

5. Konsistenz

5.1. Konsistenzprüfungen

Weiterhin spielt die Artikelstruktur eine große Rolle bei der Konsistenzprüfung, die man gerne dem Computer überlassen möchte, da sie zeitraubend und fehleranfällig ist und nicht als besonders kreative Tätigkeit gelten kann. Hier können z.B. folgende Angaben überprüft werden:

1. Sind alle obligatorischen Angaben gemacht?
   Z.B.: Wurden die Angaben zur Grammatik nicht vergessen?

2. (a) Gibt es innerhalb des Artikels einen Verweis auf eine Artikelposition, die nicht vorhanden ist?

   (b) Gibt es im Lexikon einen Verweis auf einen Artikel der nicht vorhanden ist?

3. Kommen die Verweise auf andere Artikel auch nicht im Vokabular der Beschreibungs- bzw. Definitionsposition des Artikels vor?
   (Dadurch soll verhindert werden, daß ein Lemma durch ein noch zu erklärendes Lemma beschrieben wird.)

4. Im Beschreibungsteil eines Artikels wird, um den Schreibaufwand für den Lexikographen niedrig zu halten, i.d.R. eine Abkürzung für das Lemma (Stichwort) benutzt. Hier wäre eine Überprüfung anhand eines lemmatisierten Registers sinnvoll.

5. Belege aus der Textdatenbank sollten automatisch an die richtige Position des Artikels gebracht werden. Hier könnte eine Überprüfung des Standards hilfreich sein, um auf diese Weise Beispiele und Zitate in einer homogenen Form zu präsentieren. Darüber hinaus ist eventuell eine Hervorhebung des Lemmas, etwa in Form von *Kursiv-* und/oder **Fett**schrift erwünscht.

6. Analog zu Punkt 5. sollte die Zitierung von Literatur einheitlich erfolgen. Dazu ist an einen Anschluß der Bibliographischen Datenbank BIDA an die Lexikographische Datenbank gedacht. Auch hier sollte es möglich sein, die in BIDA gespeicherten Daten in standardisierter Form in die Literaturangabe eines Wörterbuchartikels kopieren zu können, bzw. falls die Angaben dazu in BIDA noch fehlen, diese aufzunehmen.

7. Eine etwas speziellere Konsistenzprüfung könnte die folgende sein. Um die Position Etymologie zu vereinheitlichen, kann etwa daran gedacht werden, standardmäßig eine feste Folge von Abstammungen zuzulassen, z.B. griech.-lat., griech.-arab.-frz., lat.-ital., ahd.-mhd. u.dgl. Damit würden Folgen der Art lat.-griech. oder mhd.-ahd. zurückgewiesen. Um das Konsistenzprüfungsverfahren etwas flexibler zu gestalten, könnte man daran denken, diese Prüfung interaktiv durchzuführen, wobei bei jeder Abweichung vom Standard dem Bearbeiter die Möglichkeiten gegeben wird, diese trotzdem gelten zu lassen.

Als weitere Konsistenzprüfungsverfahren im weiteren Sinne des Wortes, weil sie, ähnlich wie eventuell Punkt 7., nicht vollautomatisch sondern

nur unterstützend bei der Artikelüberprüfung eingesetzt werden können, kann man an folgende Verfahren denken:

8. Erstellung eines Wortformen-Registers des Beschreibungsteils eines Artikels und Vergleich mit der Lemmakandidatenliste (Stichwortliste). Anhand von Übereinstimmungen kann der Lexikograph dann erkennen, ob das Vokabular des Beschreibungsteils eventuell geändert werden muß (Überarbeitung dieser Artikelposition).

9. Um festzustellen, ob zwischen den einzelnen Bearbeitern des Lexikons ein möglicherweise zu großer Unterschied bezüglich der Artikelstrukturierung vorhanden ist, kann man an eine graphische Darstellung der Artikelstruktur denken. Diese kann dem Bearbeiter auch als symbolisches Stenogramm seines Artikels dienen und vielleicht leichter auf Unvollständigkeiten hinweisen.

10. Analog zu Punkt 9. tritt bei der Bearbeitung eines Lexikons durch viele Mitarbeiter der Wunsch eines einzelnen auf, zu wissen, welche Artikel anderer Bearbeiter in einem gewissen Bezug zu dem gerade zu bearbeitenden Artikel stehen. Beispielsweise ist es hier wichtig, bestimmte Positionen auf Einheitlichkeit zu untersuchen. Da oft jedoch der reine Formalismus zur Konsistenzprüfung nicht reicht, sollten die entsprechenden Positionen dem Lexikographen zur Durchsicht und Überprüfung dargeboten werden. Um dies an einem Beispiel zu erläutern: Der Bearbeiter des politischen Wortschatzes (vgl. *Revanchismus, Pazifismus, Kommunismus,...)* interessiert sich, welche etymologischen Angaben in anderen Artikeln, in denen das Morphem *-ismus* vorkommt, gemacht wurden. Hier sieht man zum Beispiel sehr deutlich, wie Konsistenzprüfung und der Zugriff auf einzelne/mehrere Artikel(positionen) im Zusammenhang stehen.

5.2. Konsistenz bei Änderungen

Das Problem der Konsistenz eines Wörterbuch(s)(artikels) tritt in noch größerem Umfang bei Änderungen von Artikeln auf.

Die Konsistenz der Makrostruktur wird vor allen Dingen durch Löschen oder Neuaufnahme eines Artikels beeinflußt, da u.U. das Verweissystem erheblich verändert wird. Im ersten Fall müssen i.d.R. alle Artikel, die sich in der einen oder anderen Weise auf ihn beziehen, überarbeitet werden.

Häufiger als die Makrostruktur wird die Mikrostruktur bei Änderungen in Mitleidenschaft gezogen, da i.d.R. das Überarbeiten eines Artikels die Stellung der Artikelpositionen und die Verweise auf sie ändert. So müssen Verweise, die von einer Artikelposition auf eine andere zeigen, bei Umstellung der Unterartikel geändert werden. Darüber hinaus besteht zwischen den einzelnen Artikelpositionen auf den unterschiedlichen Hierarchieebenen ein Zusammenhang. Es gibt z.B. folgende Möglichkeiten:

1. Die Angaben einer Artikelposition (z.B. Etymologie) auf einer höheren Ebene gelten auch für die darunter liegenden (z.B. wenn keine andere Angabe gemacht wurde), d.h. sie 'vererben' sich.
2. Wird auch auf der niedrigeren Ebene eine Angabe gemacht, so sind zwei Fälle denkbar:
   (a) die Angabe der hierarchisch höher stehenden Ebene gilt nicht mehr
   (b) die Angabe der hierarchisch höher stehenden Ebene wird ergänzt

Dabei ist vor allem zu bedenken, daß für verschiedene Artikelpositionen verschiedene Fälle auftreten können, z.B. kann für Angaben zur Grammatik eine andere Regel gelten als für die Position der Beschreibung. Sollte die Anordnung der Unterartikel geändert werden, etwa der Art, daß Unterartikel III an die erste Stelle kommen soll, die Unterartikel I nach II und II nach III, so müßten diese Abhängigkeiten mit berücksichtigt werden. Dabei muß man deutlich hervorheben, daß von seiten des Computers keine Sortierung mit Berücksichtigung der Semantik erfolgen kann, jedoch eine Sortierung, die man als 'kontextabhängig' bezeichnen könnte. Bedingung dafür ist, daß sich die Artikelpositionen klar definieren und markieren lassen, daß Angaben über die 'Vererbbarkeit' von Eigenschaften gemacht werden können und die Bedeutung (Semantik) der einzelnen Unterartikel unabhängig voneinander ist (abgesehen von den Verweisen aufeinander).

Sind diese Voraussetzungen erfüllt, kann nach eventuell interaktiver Angabe der 'Vererbungsregeln' für den Artikel durch den Lexikographen die Sortierung durchgeführt werden. Man kann etwa folgendermaßen vorgehen: Der Artikel wird durch ein semantisches Netz dargestellt, die einzelnen Unterartikel werden dabei durch Knoten dargestellt, die durch Kanten hierarchisch verbunden sind. Diesen Knoten wird jeweils ein anderer Typ von Knoten zugeordnet, der die Artikelpositionen repräsentiert. Diesen Positionen wird eine Regel zugeordnet, die die Form der Vererbung wiedergibt. Bevor die Unterartikel sortiert werden können, muß zuerst die ge-

samte Information, die sich in den Artikelpositionen befindet (also vor allem die durch die Vererbungsregeln hervorgerufene implizite Information) gesammelt werden (Explizitmachen). Danach ist die Sortierung einfach. Den sortierten Artikel selbst erhält man jedoch erst, wenn man anhand der angegebenen Regeln die redundante Information wieder reduziert.

6. Realisierung

Die oben genannten Beispiele haben deutlich gemacht, daß eine Abbildung von Artikelstrukturen auf Datenstrukturen nicht ohne Schwierigkeiten stattfinden kann, und daß man hier nicht einfach herkömmliche Modellierungskonzepte anwenden kann. Im folgenden soll versucht werden, die oben schon erwähnten Benutzerwünsche (in der Datenbank-Notation: externes Schema genannt) zu einem einheitlichen Konzept zusammenzufassen (konzeptuelles Schema) und anhand dessen ein mögliches Datenmodell zu erstellen (internes Schema). Es soll dabei lediglich eine mögliche Realisierung skizziert werden, wobei offen bleiben soll, ob und wenn ja welches kommerzielle Datenbank-System zum Einsatz kommt und welche Programmiersprachen benutzt werden.

Es soll des weiteren in diesem Zusammenhang genügen, das System zu skizzieren und eine kurze Begründung für die einzelnen Komponenten zu geben.

Um dem Lexikographen möglichst viel Flexibilität bezüglich der Strukturierung seiner Artikel zu erlauben, ist daran gedacht, statt der oben geschilderten Datenmodelle, die, wie wir gesehen haben, alle ein zu enges Korsett für die Repräsentation der Artikelstrukturen darstellen, die Artikel zunächst als ganz normalen Text über einen Editor in den Computer einzugeben. Dies geschieht in der Arbeitsdatei. Zur Markierung der Artikelstruktur wird eine Codierung definiert, die aus Metazeichen besteht und Hierarchieebene, Artikelposition und Verweise kennzeichnet. Anhand von Regeln können verschiedene Programme die Konsistenz des Artikels prüfen und seine Struktur in einer Reihe von Dateien, die zusammengenommen die oben genannte Ergebnisdatenbank darstellen und nach dem Konzept der relationalen Datenbanken angelegt sind, adäquat abspeichern. Dadurch ist gewährleistet, daß der Bearbeiter zur Strukturierung ein Mittel zur Verfügung hat, das seiner üblichen Formulierung entgegenkommt und das flexibel ist.

Besehen wir uns als nächstes eine mögliche Form der Realisierung der Ergebnisdatenbank (die dazugehörigen Graphiken befinden sich im Anhang):

Die Datei I enthält zu jedem Lemma eine Nummer, die zu Adressierungszwecken verwendet wird, und ein Feld mit der Angabe, ob zu diesem Lemma schon ein Artikel aufgenommen wurde. (Die Numerierung ist von Vorteil, da sie Platz spart: die Länge der Lemmata streut zu sehr und die Adresse wird in vielen Dateien benutzt!) Diese Datei kann eventuell in der 'Wortdatenbank' integriert sein und dient - abgesehen von der Adressierung - für einen schnellen Überblick.

Datei II enthält umgekehrt zu jeder Lemma-Nummer das dazugehörige Lemma (Anwendung s. Datei VII).

Datei III enthält zu jeder Lemma-Nummer eine Kennzeichnung der Artikelstruktur. Diese wird entweder am Anfang vom Lexikographen festgelegt oder zu Beginn eines jeden Artikels mit Hilfe eines Metazeichens markiert und vom Prüfprogramm in die Datei III eingetragen. Verschiedene Strukturen wären beispielsweise onomasiologische versus semasiologische Artikel (bzw. Einzel- versus Rahmenartikel, vgl. Vortrag Gerhard Strauß).

Datei IV enthält Angaben zu den Artikelstrukturen. Hier soll festgelegt werden, welche Artikelpositionen obligatorisch und welche fakultativ sind, wobei u.U. eine Unterscheidung zwischen Vorspann und Unterartikel nötig sein kann (zusätzliches Feld), und vor welcher anderen Artikelposition diese unbedingt im Artikel auftreten muß, z.B. 1 vor 2, 2 vor 3 u.dgl.

Datei V dient dazu, der Codierung der Artikelpositionen die ausführliche Bezeichnung der Position (z.B. 'Etymologie') und den Namen der Datei, in der sich die dazugehörigen Textbausteine befinden, zuzuordnen.

Die Datei VI enthält die Mikrostruktur eines Artikels. Der Schlüssel dieser Datei setzt sich zusammen aus der Nummer aus Datei I und einer Nummer für den jeweiligen Unterartikel, beginnend mit der Zahl Null für den Vorspann. Die weiteren Felder bezeichnen die Nummer für den Unterartikel auf gleicher Stufe und den auf der nächst niedrigeren Ebene. Diese Angaben zusammen mit der Nummer des Lemmas erlauben die Adressierung der weiteren Unterartikel. Man beachte in diesem Zusammenhang, daß die Anzahl der Hierarchie-Stufen schlimmstenfalls durch die Länge der Felder für die Unterartikel-Adressierung begrenzt ist (2 Byte binär: max. Anzahl der U-Artikel ≈ 65.000!). Die weiteren Felder dieser Datei enthalten Angaben zu

den Artikelpositionen, die der Unterartikel enthält. Eine feste Struktur in Form einer fixen Anzahl von Feldern (bzw. Artikelpositionen) scheint hier nicht weiter hinderlich zu sein, da eine zu starke Unterteilung in Teile von Artikelpositionen bald an die 'Satzgrenzen' im linguistischen Sinne stößt oder schon nicht mehr als strukturiert angesehen werden kann. Die Angaben zu den Artikelpositionen bestehen aus einer Kennzeichnung für die Position und einer für die Reihenfolge. Letztere ist nötig, um alternierendes und geschachteltes Auftreten von Artikelpositionen zu erlauben und ist Bestandteil des Schlüssels der eigentlichen 'Textdateien' (Datei VII).

Die Bezeichnung Datei VII steht für eine Gruppe von Dateien, die jeweils eine Artikelposition enthalten. Der Schlüssel setzt sich zusammen aus der Lemma-Nummer, der Unterartikel-Nummer und der Kennzeichnung für die Unterposition. Zusammen mit II erlaubt diese Art der Realisierung das schnellere Durchsuchen einzelner Artikelpositionen nach Besonderheiten und die Angabe des Artikels, in der diese auftraten.

Die bisherige Modellierung berücksichtigt die Hierarchie, einen schnellen Datenzugriff nach verschiedenen Aspekten, zum Teil die Darstellung des Layout und ermöglicht eine Konsistenzprüfung. Die nicht-lineare Struktur eines Wörterbuches, die durch die Verweise zustande kommt, blieb bisher unberücksichtigt und muß nach unserem Modell mit Hilfe weiterer Dateien dargestellt werden. Diese sog. invertierten Listen können nach verschiedenen Kriterien angelegt werden. Naheliegende Möglichkeiten für eine Klassifizierung sind:

(a) Verweisart:

  i) Synonym/Antonym
  ii) Hyperonym/Hyponym
  iii) Kohyponym

(b) Artikelposition, in der der Verweis auftritt

(c) Ausgangs- oder Zielliste;

  d.h. entweder alle Verweise, die in einem Artikel bzw. in einer Artikelposition auftreten oder alle Lemmata, von denen aus auf den entsprechenden Artikel verwiesen wird

Würde man alle Kombinationen aus (a) bis (c) berücksichtigen, erhielte man eine fast unüberschaubare Ansammlung von Dateien. In diesem Zusammenhang ist jedoch nicht nur der dadurch benötigte Speicherplatz von Interesse, sondern vor allem die dadurch entstehende Komplexität, was die Verwaltung, vor allen Dingen die Änderung von Daten bzw. Strukturen, betrifft. Hier muß vor der Implementierung des Systems klar überlegt werden, welche Zugriffsmöglichkeiten auch im Zusammenhang mit der Prüfung der Konsistenz unbedingt nötig sind, und auf welche man getrost verzichten kann.

Auf das Thema 'Verweise in Lexikographischen Datenbanken' wird hier nicht näher eingegangen, da es Gegenstand des anschließenden Vortrags ist.

7. Zusammenfassung

Wir haben nun folgendes gesehen:

1. Eine Lexikographische Datenbank besteht i.d.R. aus einer Vielzahl von Komponenten (Text-Datenbank, Bibliographie-Datenbank, einer (oder mehreren) Wortliste(n), Arbeitsdateien und Ergebnis-Datenbank).
2. Eine Lexikographische Datenbank muß durch eine Reihe von Programmen unterstützt werden (z.B. komplexe Zugriffsmöglichkeiten, (halb)automatische Aufnahme von Belegen und bibliographischen Angaben an die richtige Artikelposition, Prüfung von makrostruktureller Kohärenz und mikrostruktureller Konsistenz u.dgl.).
3. Die Artikelstruktur bestimmt im wesentlichen die Auswahl des Datenmodells. Dabei stellen kommerzielle Datenbanksysteme i.d.R. zur Realisierung einer Lexikographischen Datenbank keine adäquaten Mittel der Datenstrukturierung bereit.
4. Die Artikelstruktur (und damit auch die Datenstruktur) kann zumindest teilweise im Zusammenhang mit dem Layout des Lexikons gesehen werden.
5. Durch die Wahl einer Artikelstruktur werden im wesentlichen die Zugriffs- bzw. Abfragemöglichkeiten (zumindest die optimalen) festgelegt.
6. Die Struktur eines Artikels legt auch die Prüfverfahren fest.

Literatur

Heß, Klaus / Brustkern, Jan / Lenders, Winfried (1983): Maschinenlesbare deutsche Wörterbücher. Tübingen 1983.

Schlageter, Gunter / Stucky, Wolffried (1983): Datenbanksysteme: Konzepte und Modelle. Stuttgart 1983.

Schwarze, Christoph / Wunderlich, Dieter (Hrsg.)(1985): Handbuch der Lexikologie. Königstein/Ts. 1985.

Wahrig, Gerhard (1973): Anleitung zur grammatisch-semantischen Beschreibung lexikalischer Einheiten. Tübingen 1973.

Anhang

(I)

Lemma	L-Nr	V
Metapher	333	J

(II)

L-Nr	Lemma
333	Metapher

(III)

L-Nr	A-Str
333	R

L-Nr = Lemma-Nummer
V = Vorhanden (bezieht sich auf Artikel)
A-Str = Artikel-Struktur (in codierter Form)

(IV)

A-Str	$A\text{-}Pos_1$	Sta_1	$A\text{-}Pos_2$	Sta_2	...
R	1	o	2	o	...

A-Str = Artikel-Struktur (in codierter Form)
$A\text{-}Pos_1$ = Artikel-Position (bzw. Textbaustein des Artikels)
Sta_1 = Status der Artikel-Position (obligatorisch oder fakultativ)

V

A-Pos	A-Pos-N	A-Pos-D
1	Lemma	LEM
2	Wortart	WORTA
...		
7	sp. Gebr.	SPEGE

A-Pos = Artikel-Position (bzw. Textbaustein)
A-Pos-N = Name der Artikel-Position
A-Pos-D = Name der Datei, in der sich die Texte der dazugehörigen Artikel-Position befinden

VI

L-Nr	U-Art	V-GS	V-NS	$A\text{-}Pos_1$	$U\text{-}Pos_1$	$A\text{-}Pos_2$	...
333	0	0	0	1	0	2	...

L-Nr = Lemma-Nummer
U-Art = Unter-Artikel (Nummer der Hierarchie-Ebene)
V-GS = Verweis auf Unter-Artikel auf gleicher Ebene
V-NS = Verweis auf Unter-Artikel auf nächst niedrigeren Ebene
$A\text{-}Pos_1$ = Nummer der Artikel-Position, die sich an 1. Stelle befindet
$U\text{-}Pos_1$ = Nummer der Unter-Position

VII

SPEGE

L-Nr	U-Art	U-Pos	Text
333	0	1	Nehmen wir die beiden folgenden Äußerungen...
333	0	2	Diese Präzisierung bringt...
333	0	3	Die metaphorische Verwendung...

L-Nr = Lemma-Nummer
U-Art = Nummer des Unter-Artikels
U-Pos = Nummer der Unter-Position